Journal of Computers Vol. 36 No. 3, June 2025, pp. 273-286
https://doi.org/10.63367/199115992025063603018

Research on Intelligent Detection and Prevention Mechanism of Malicious Traffic in the Internet of Things Based on Machine Learning

Shao-Ting Liu[1,2,3], Jun-Xia Zhang[1], Meng-Ying Yang[1],
Xin-Hong Hu[1*], and Zhan-Feng Yang[4]

[1] Hebei Institute of Mechanical and Electrical Technology,
Xingtai City 054000, Hebei Province, China
{stliu_hbjd, junxia9987, mengying20180107, xinhong5817}@163.com

[2] Intelligent Sensor Network Application Technology Innovation Center,
Xingtai City 054000, Hebei Province, China

[3] Xingtai City Smart Manufacturing and Digital Twin Technology Innovation Center,
Xingtai City 054000, Hebei Province, China

[4] State Grid Hebei Electric Power Co., Ltd. Nangong Power Supply Branch,
Xingtai City 055750, Hebei Province, China

my992844593@163.com

Received 20 May 2025; Revised 5 June 2025; Accepted 9 June 2025

Abstract. The increasing integration of Internet of Things (IoT) devices into critical infrastructure has brought unprecedented connectivity and significant information security challenges such as malicious traffic and network intrusions. This paper applies machine learning (ML) techniques to intelligently detect and prevent such malicious traffic in IoT environments. The paper explains in detail how ML techniques such as supervised learning, unsupervised learning, deep learning, and federated learning can be used to enhance the security posture of IoT by detecting malicious patterns in real time, adaptively, and accurately. Special emphasis is placed on the malicious traffic categories that are unique to IoT, the shortcomings of traditional intrusion detection systems, and how intelligent ML-based approaches can overcome these issues through autonomous feature learning and behavioral analysis. Key challenges such as data imbalance, weak generalization capabilities, real-time processing requirements, and model vulnerability to adversarial attacks are discussed in detail. The paper concludes by pointing out future research avenues for efficient, scalable, and explainable machine learning-based security solutions for IoT, making this review a comprehensive roadmap for advancing intelligent information security to address impending cyber threats.

Keywords: intelligent detection, malicious traffic, internet of things, machine learning, network security

1 Introduction

The large-scale deployment of Internet of Things (IoT) devices in application areas such as smart life and healthcare, industrial automation, and critical infrastructure has transformed the digital world with unprecedented levels of connectivity and automation. However, as the scale of IoT devices continues to grow, security threats are also increasing, mainly due to the heterogeneity, scalability, and resource-constrained nature of IoT devices, which often lack strong built-in security mechanisms. As a result, IoT networks have become a prime target for various cyberattacks, especially malicious traffic attacks, including botnet communications, command injection attacks, and protocol abuse or anomalies. Such attacks severely undermine the basic principles of network security (availability, integrity, and confidentiality) as they lead to unauthorized access to information, disruption of device functionality, and cause large-scale denial of service attacks. Traditional rule-based or signature-based security solutions have proven to be inadequate as they lack sufficient flexibility and cannot identify unknown threats. Therefore, we need intelligent, real-time, and adaptive detection technologies that can cope with dynamic threat environments. Machine learning (ML) is a promising solution that uses data-driven models to learn com-

* Corresponding Author

plex malicious activity patterns from network traffic and generalize them to the detection of new types of attacks. This paper presents a comprehensive overview of the state-of-the-art in machine learning-based malicious IoT traffic detection and related technologies, with a focus on detection, anomaly classification, and deep learning techniques. We review existing methods based on their methodological design, compare their detection performance on common datasets, and identify challenges in current research, such as data bias, interpretability, and real-time operational constraints. Although numerous machine learning-based countermeasures have shown promise in IoT attack detection, there are still some important gaps to be filled, especially in dealing with the dynamic nature of IoT environments, the high computational cost of some machine learning models, and how to achieve interpretability in the decision-making process of security practitioners. In addition, the paper highlights the lack of common evaluation metrics across different IoT networks and the lack of large labeled datasets for training robust models as prominent obstacles that hinder research progress.

2 Malicious Traffic in IoT Networks

Malicious traffic in IoT networks is difficult to distinguish from legitimate traffic due to encryption, device diversity, and complex behavior. The study highlights key metrics such as flow, timing, protocol anomalies, and source diversity. Advanced techniques such as deep learning, entropy-based feature selection, and attention mechanisms improve detection accuracy. Attack types include DDoS, spoofing, and botnets, while mitigation strategies leverage SDN, blockchain, and AI. Flow-based features are critical, especially when detecting encrypted threats. In summary, a multi-layered, behavior-driven approach is essential to protect the evolving IoT environment.

2.1 Characteristics of Malicious Traffic in IoT

The IoT security landscape is particularly challenging due to the difficulty in distinguishing legitimate traffic from malicious traffic. As IoT networks continue to grow in both size and complexity, the capability to detect and classify traffic effectively is increasingly important to their integrity and operation. A number of researchers have suggested various approaches to characterize malicious traffic in IoT environments with an emphasis on the prominent features of volume, timing, source diversity, and protocol anomalies of traffic. Pinheiro [1] identified the challenge of external attackers in detecting active IoT devices in encrypted traffic, which prefers to exploit traffic and protocol anomalies. Zhu [2] proposed a CMTSNN model for multi-classification of malicious and encrypted IoT traffic, with a priority on the metrics accuracy, precision, recall, and false positive rate, which are critical to distinguish between normal and malicious traffic. Let *A* denote the set of accurate classifications, *P* the set of predictions, T_{true} the true traffic classification, and T_{pred} the predicted traffic. The evaluation metrics are then defined as:

$$\begin{cases} Accuracy = \dfrac{\left|A \cap T_{ture}\right|}{\left|T_{ture}\right|} \\ Precision = \dfrac{\left|A \cap T_{pred}\right|}{\left|T_{pred}\right|} \\ Recall = \dfrac{\left|A \cap T_{true}\right|}{\left|T_{true}\right|} \\ FalsePositiveRate = \dfrac{\left|T_{pred} - A \cap T_{true}\right|}{\left|T_{pred}\right|} \end{cases} \tag{1}$$

These metrics are essential for achieving an effective classification of normal and malicious traffic in IoT networks. Diwa [3] introduced a novel feature selection technique for IoT traffic detection, with consideration to feature entropy estimation (FEE), which helps to accurately identify anomalies by considering traffic source diversity and protocol misuse. Similarly, Kawai [4] demonstrated that the identity of IoT devices can be revealed by communication traffic patterns, which confirms the importance of unique traffic patterns in IoT security. In addition, Islam [5] proposed a stacked ensemble (SE) learning method for the classification of malicious activities, which investigates traffic features such as flow, time, and protocol anomalies. Koroniotis [6] targeted the identification of botnets in IoT networks, more specifically botnets that exploit various traffic sources and large amounts of traffic from infected devices. Liao and Guan [7] proposed a multi-scale convolutional feature fusion network based on an attention mechanism, which was an excellent improvement in protocol anomaly detection and traffic variation detection based on deep learning techniques. Let the traffic features be represented by F_i for different scales, and the fused feature set can be expressed as:

$$F_{fused} = \sum_{i=1}^{n} \alpha_i F_i \tag{2}$$

Where is α_i the attention weight applied to the features F_i, which allows the network to focus on the most relevant features for traffic anomaly detection.

Together, these works showcase the multifaceted nature of malicious traffic comprehension and identification in IoT networks. Spanning the application of machine learning to the analysis of traffic patterns, these works emphasize the indispensable contribution of traffic features to the continual protection of IoT environments.

2.2 Types and Sources of Attacks Generating Malicious Traffic

The cyber threat landscape against the IoT and smart devices is similarly becoming more heterogeneous and sophisticated, and it poses a significant threat to the security of connected systems. Common threats such as distributed denial of service (DDoS) attacks, spoofing, packet flooding, botnet propagation, man-in-the-middle attacks, and ransomware can generate malicious traffic that can have a substantial impact on the functionality and security of IoT devices and the IoT ecosystem in general. To combat these attacks, current research has attempted to provide a variety of new strategies. One promising strategy is to use software-defined networking (SDN) for monitoring devices to be compliant with manufacturer usage description (MUD) behavior profiles and combining this with machine learning techniques to monitor volumetric attacks such as DoS, reflection flooding, and ARP spoofing [8]. Let the total volume of network traffic be represented by V_{total}, and the malicious traffic volume by V_{mal}. The detection of volumetric attacks can then be expressed as:

$$V_{mal} = f(T_{dos}, T_{reflection}, T_{arp\_spoof}) \tag{3}$$

Where T_{dos} , $T_{reflection}$, T_{arp_spoof} , represent the traffic corresponding to different volumetric attacks (DoS, reflection flooding, and ARP spoofing). The approach leverages the programmability of SDN to facilitate real-time monitoring and effective threat management, demonstrating its feasibility in mitigating volumetric attacks. The ransomware threat in IoT systems has also received a lot of attention, with numerous studies investigating high-tech approaches to identify ransomware attacks, such as artificial intelligence (AI), blockchain, and SDN [9]. Such research highlights the importance of adopting a multi-dimensioned approach to effectively fight the richness and dynamic nature of cyber attacks. Collective strategies leverage smart contracts for enabling mitigation of DDoS in the scope of SDN domains to provide flexibility, efficiency, and security in distributed exchange of threat information [10]. Additionally, deep learning has been applied in detecting DDoS attacks, wherein neural network frameworks identify complex application-layer DDoS activities [11]. Blockchain technology is also under study to support data security in IoT usage, especially in intelligent healthcare, owing to its immutability and transparency [12]. With more sophisticated attacks taking place, it is imperative that research continues to develop smarter detection and mitigation solutions that utilize smarter technologies like deep learning, blockchain, and SDN to secure IoT networks.

2.3 Behavioral Patterns and Traffic-Level Indicators

In network security, the identification of malicious activity in network traffic is now an essential challenge, especially with increasing amounts and complexities of IoT traffic. Scientists are resorting to flow-based features such as packet size, flow duration, burstiness, and entropy for enhancing detection systems. Ali’s [13] work added much value to this field by proposing a multi-task deep learning (DL) model for classifying traffic as malicious or benign, and the type of malware. This research highlights the significance of traffic analysis based on behavior in IoT systems and demonstrates that DL models can be used efficiently to identify malware based on traffic features. Another novel method is introduced by Fu [14], who introduced HyperVision, an unsupervised graph learning method to inspect encrypted malicious traffic.

Their method builds graph features from flow interaction patterns and can detect unknown attacks without labeled datasets, which is highly crucial to detect encrypted traffic behaviors typically overlooked by traditional detection methods. While Feng [15] were focusing on social robots, they also proposed a new approach to detect malware based on traffic features. The importance of behavioral pattern analysis for network security is also emphasized. Their BotShape approach showed that the same flow-based behavioral features can be utilized to identify malicious activity in network traffic. Papadogiannaki and Ioannidis [16] had carried out a critical review of anomaly detection methods from encrypted traffic applying AI, further highlighting the issue and innovation in detecting anomalies of encrypted data streams, and Imtiaz [17] offered further insights about how flow-based features can be utilized in real-time as well as encrypted network environments. Overall, these studies illustrate the potential consensus that flow-based features play a critical role in detection of malicious activity in complex network settings. Deep learning models, graph learning without supervision, and AI-based anomaly detection techniques are improving detection accuracy, laying the basis for future research focused on further improving threat detection in complex networks.

3 Evaluation Metrics and Benchmark Datasets

In the field of IoT security, performance metrics such as accuracy, precision, recall, and F1 score are critical for evaluating and optimizing machine learning models. While datasets such as BoT-IoT and CICIDS2017 support intrusion detection, issues such as class imbalance and static data remain challenging. Advanced models can improve detection capabilities, but usually at an increased computational cost, highlighting the need to strike a balance between accuracy, efficiency, and practical deployment capabilities.

3.1 Performance Metrics for Security Effectiveness

With ML and DL techniques merging together, the field of cybersecurity has also evolved to a large extent to tackle more advanced cyber attacks and anomalies. The increasing complexity of network infrastructure, especially with the proliferation of IoT devices and cloud computing, highlights the need for stringent security mechanisms at play to quantify the performance of these intelligent security systems. Important performance metrics such as accuracy, detection rate, false alarm rate, precision, recall, F1 score, and AUC-ROC not only are required to quantify the performance of cybersecurity solutions, but also to fine-tune the detection algorithms in Table 1. In the field of IoT security, Rahim [18] studied the use of Logit-Boosted CNN models to enhance the security of smart homes, especially in the areas of anomaly detection and facial recognition. Although their work did not directly relate to performance measures, precision and recall would be one area of measuring the effectiveness of such models in differentiating between real users and intruders. Shagari [19] examined Logit-Boosted CNN models as part of fortifying smart home security, especially anomaly detection and face recognition. While their research does not specifically mention performance measures, precision and recall would be used in order to identify the accuracy of such models in being able to separate legitimate users from attackers, similarly highlighting the detection accuracy and false positive rate within their hybrid logistic regression (LR) model used for warding off SQL injection attacks in web applications and identifying such metrics as necessary for upholding system integrity as well as user trust. In speaker identification, the use of feature fusion and ML techniques to further optimize model performance further emphasizes metrics such as accuracy, precision, and recall, especially in challenging situations such as voice variation and ambient noise [20]. For network intrusion detection, Çoşkun and Çetin [21] contrasted the performance of CatBoost classifier in distinguishing between malicious and benign traffic with precision, recall, and F1 score identified as important metrics to accurately classify normal and abnormal network behavior. Finally, Liu [22] addressed ensemble learning for spam filtering on obtaining a tradeoff between high recall and low false positives, which is an important consideration for user-end applications. Together, these studies demonstrate the essential role of performance metrics in cybersecurity for informing algorithm improvement, model choice, and security standardization to counter the evolving cyber threats.

Table 1. Summary table of the study

Study	Focus area	Key performance metrics	Insights
Rahim et al. [18]	IoT security (Smart homes)	Precision, Recall	Focused on enhancing security via Logit-Boosted CNN models, with precision and recall metrics being critical for differentiating between real users and intruders.
Shagari et al. [19]	Smart home security (Anomaly detection & Face recognition)	Precision, Recall	Emphasized the importance of precision and recall in identifying attackers vs. legitimate users, and the need for accuracy in preventing attacks like SQL injection.
Jahangir et al. [20]	Voice recognition	Accuracy, Precision, Recall	Discussed the use of feature fusion and ML techniques in voice identification under challenging conditions, with key metrics to assess model effectiveness.
Çoşkun and Çetin [21]	Network intrusion detection	Precision, Recall, F1 Score	Contrasted CatBoost classifier performance for identifying malicious vs. benign network traffic, highlighting precision, recall, and F1 score for classification accuracy.
Liu et al. [22]	Spam filtering	Recall, False Positives	Addressed the tradeoff between high recall and low false positives in ensemble learning for spam filtering, crucial for maintaining user experience.

The changing IoT security landscape calls for reliable datasets to support network traffic analysis in ensuring efficient intrusion detection and security. Datasets such as TON_IoT, BoT-IoT, CICIDS2017, and N-BaIoT are now at the core of benchmarking a variety of intrusion detection systems (IDS) and network analytical platforms. Different challenges and possibilities face each dataset in the development of sophisticated security mechanisms for IoT networks. For example, the Bot-IoT [23] dataset provides a large dataset that includes both normal and synthetic IoT network traffic, which provides a rich source of botnet behavior to be detected in IoT networks. Although it is beneficial, issues such as class imbalance (benign instances overwhelmingly surpass malicious instances) have been surmounted, and Wheelus et al. [24] used synthetic data creation techniques such as SMOTE to balance the data set and therefore improve the performance of machine learning algorithms. Similarly, the CICIDS2017 data set has received extensive attention due to its comprehensive nature of network traffic. Studies comparing machine learning models on this dataset have indicated the effectiveness of supervised and unsupervised models in anomaly detection, which can be applied to IDS optimization. To address the class imbalance problem in CICIDS2017, Barkah et al. [25] suggested a generative model to improve the minority class representation, thereby improving the precision and recall values in IDS application. In the case of N-BaIoT, machine learning techniques such as logistic regression and artificial neural networks have been utilized in the identification of anomalies, with it being advised that there should be a compromise between feature engineering and model complexity in order to achieve high classification accuracy.

The TON_IoT dataset is dedicated to IIoT and IoT device telemetry data and is pivotal in encouraging the growth of IDS through the support of federated learning and hybrid machine learning algorithm usage. This supports the development of IDS frameworks that can handle the diverse and large amounts of data produced by IoT devices. Briefly, while datasets such as TON_IoT, BoT-IoT, CICIDS2017, and N-BaIoT are required to advance IoT security research, class imbalance and dynamic nature of network traffic are still issues. Future research can include creating more dynamic datasets in combination with federated learning, or exploring GANs to enhance existing datasets, such that IDS models are learned on data which actually reflects the changing threat landscape in IoT networks.

3.2 Comparative Summary of Reviewed Approaches

Recent research in the field of ML focuses on maximizing the accuracy of the model while ensuring computational efficiency, flexibility, and deployability across various applications. Significant areas where ML models have great potential are power load forecasting, DDoS attack detection in IoT sensor systems, atmospheric chemistry modeling, forest fire detection, malicious URL identification, and plant water stress sensing. These applications highlight the need for robust ML models with the ability to strike a balance between high detection accuracy, minimal computational cost, and deployability in real-world applications in Fig. 1.

Rahman [26] demonstrated how an ML model that integrates big data analytics with deep learning better predicted power demand than the traditional method more precisely and in a more adaptable way, though at greater computational expense when using more sophisticated models. Similarly, in the banking domain, discovery of DDoS attacks was used to demonstrate the robustness of SVMs compared to other models such as KNN and random forests (RF), but computational cost and usability during deployment were not experimented. In computational modeling of atmospheric chemistry, Keller [27] proposed a way to compromise computational cost without compromising precision, illustrating that deployability and flexibility are still second thoughts in models of such intricate systems. Sathishkumar [28] demonstrated increased accuracy in a forest fire alarm system using CNNs but at the expense of computational complexity, which indicates the trade-offs incurred in using advanced deep learning models in actual implementations. In other domains, such as malicious URL detection and water stress detection, ML model application has consistently enhanced detection precision but the challenges of computational cost, flexibility, and real-world deployment remain prominent. In brief, while ML models have made much improvement across disciplines, achieving an equilibrium between accuracy of detection and computational efficiency on the one hand and ease of deployment on the other remains problematic. This literature survey points toward the need for further research optimizing ML models in their applicability to real-world contexts, including consideration not merely of performance metrics but also pragmatic constraints of operating these models on different operating contexts.

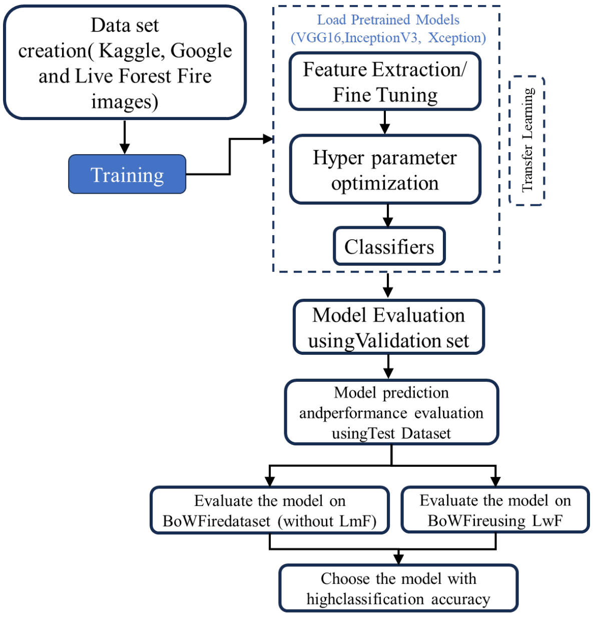
Data set creation(Kaggle, Google and Live Forest Fire images)
Training
Load Pretrained Models (VGG16,InceptionV3, Xception)
Feature Extraction/ Fine Tuning
Hyper parameter optimization
Classifiers
Transfer Learning
Model Evaluation usingValidation set
Model prediction andperformance evaluation usingTest Dataset
Evaluate the model on BoWFiredataset (without LmF)
Evaluate the model on BoWFireusing LwF
Choose the model with highclassification accuracy

Fig. 1. Proposed workflow

4 Challenges of Intelligent Traffic Detection in IoT

IoT security faces many key challenges that hinder effective model training and real-world performance. Solutions such as resampling, semi-supervised labeling, and realistic dataset generation can improve detection reliability. As IoT environments vary widely, ensuring cross-domain generalization is critical, while techniques such as domain adaptation, meta-learning, and synthesis based on generative adversarial networks can enhance model adaptability. Real-time detection is limited by the resources of IoT devices. Optimized lightweight models, edge computing, and hardware-software co-design help achieve fast and efficient inference. At the same time, machine learning models are still vulnerable to adversarial reasoning and membership inference attacks.

4.1 Data Imbalance, Labels, and Authenticity Gap

Effective detection of malicious traffic in IoT networks relies heavily on high-quality datasets. However, three persistent issues hinder progress: data imbalance, poor label quality, and lack of authenticity in simulated datasets. Data imbalance refers to the over-representation of benign traffic samples relative to malicious traffic samples, which causes the trained model to be biased towards the majority class and has difficulty identifying actual attacks. The formula to handle data imbalance using a weighted loss function is:

$$L(p,y) = -\sum_{i} w_i y_i \log(p_i) \tag{4}$$

Where p_i is the predicted probability of class, y_i is the ground truth for class, w_i is the weight associated with class i to penalize the majority class.

Inaccurate or noisy labels can interfere with learning algorithms during training, further degrading model performance. Active learning and semi-supervised methods can decrease this issue by allowing the model to query uncertain samples to be labeled by a human. The semi-supervised objective function for training with labelled and unlabelled data can be simply expressed as:

$$L_{total} = L_{supervised} + \lambda L_{unsupervised} \tag{5}$$

Where $L_{supervised}$ is the loss on labeled data, $L_{unsupervised}$ is the loss on unlabeled data, and λ is a hyperparameter controlling the weight of the unsupervised loss.

4.2 Model Generalization and Cross-Domain Transfer Capability

A major challenge facing IoT traffic detection is to ensure that models trained in one environment can perform well in other environments, which is called generalization and cross-domain transfer capability. To ensure domain-invariant feature learning, one approach is the use of adversarial training, where the model learns features that cannot be easily distinguished between domains. The minimax optimization problem for domain adaptation can be expressed as:

$$\min_{G}\max_{D} E_{x\sim P_s}[log(D(G(x)))] + E_{x\sim P_t}[log(1-D(G(x)))] \tag{6}$$

Where G is the generator network that transforms source data to target data, D is the discriminator network that tries to distinguish between source and target domains. Given labeled source domain data, the classification loss is typically defined as the cross-entropy:

$$L_{total} = L_{cls}(G,C) + \lambda \cdot L_{domain}(D,G) \tag{7}$$

In practice, this adversarial training is combined with a classification loss on the labeled source data:

$$L_{cls} = -E_{(x_s,y_s)} \sum_{k=1}^{K} 1_{[y_s=k]} logC(G(x_s))_k \tag{8}$$

Where K is the number of classes, and $C(G(x_s))_k$ is the predicted probability for class k.

4.3 Real-time Constraints and Resource Limitations

IoT devices are often severely constrained in terms of processing power, memory, and energy consumption, which complicates the deployment of complex machine learning models. Real-time detection requires low-latency, high-throughput inference capabilities, which deep learning models cannot achieve without optimization. To optimize models for real-time processing, model compression techniques like pruning or quantization are commonly applied. The objective function for pruning can be expressed as:

$$\min_{\theta} L(\theta) + \lambda \|\theta\|_0 \tag{9}$$

Where $L(\theta)$ is the model's loss function, $\|\theta\|_0$ is the L-norm of the weights, representing the number of non-zero weights, and λ is a hyperparameter controlling the sparsity of the model.

Quantization reduces the precision of weights and activations, significantly reducing memory usage and inference time. A typical quantization operation can be represented as:

$$\tilde{w} = round(\frac{w}{\Delta}) \cdot \Delta \tag{10}$$

Where w is the original weight, Δ is the quantization step size, $\tilde{w}$ is the quantized weight. Quantization-aware training introduces a loss term to penalize quantization error:

$$L_{quant} = L(f(x;\theta)) + \beta \cdot \left\|\theta - \tilde{\theta}\right\|_2^2 \tag{11}$$

Where θ is the quantized version of $\tilde{\theta}$, and β is a tuning hyperparameter.

4.4 Security and Robustness of Machine Learning Models

Machine learning models for IoT security are inherently vulnerable to attacks. Adversarial attacks can subtly modify input traffic to deceive the model, while membership inference attacks attempt to extract information about the data used during training. A robust loss function designed to defend against adversarial perturbations is the adversarial training loss, which includes adversarial samples generated by a separate model. The robust objective can be formulated as:

$$L_{adv}(x, y) = E_{x' \sim A(x)}[L(x', y)] \tag{12}$$

Where $A(x)$ generates adversarial samples for a given input x, $L\left(x^{'}, y\right)$ is the loss on the adversarial sample $x^{'}$, y is the label of the original input. Adversarial attacks exploit the sensitivity of models to carefully crafted, small perturbations. These perturbations may not affect the overall semantics of the traffic, but may significantly change the model's predictions, leading to false positives or false negatives. On the other hand, membership inference attacks exploit the model's behavior to determine whether a specific data point belongs to the training dataset. Such vulnerabilities highlight the importance of not only optimizing detection accuracy but also enforcing the confidentiality and integrity of the underlying model.

Defending against these threats requires a robust training procedure. One approach is adversarial training, which introduces perturbed samples during the model learning process to enhance resilience. Defensive distillation can also be used to smooth the decision boundaries of the model, making it more difficult for attackers to exploit subtle changes in Fig. 2. Anomaly detection wrappers (external models that monitor output anomalies) provide an additional layer of security, especially in mission-critical systems.

Another important aspect of robustness is ensuring that models remain effective even in the face of evolving threats and changing data patterns. Continuous learning techniques are essential to maintaining up-to-date detection capabilities. In addition, ensemble learning can improve robustness by compensating for the weaknesses of individual models and reducing the likelihood of simultaneous failures.

Model interpretability is equally important. In high-risk applications, decision transparency enables human analysts to trace, verify, and understand the rationale behind an alert or classification. Techniques in Explainable Artificial Intelligence (XAI), such as SHAP values, integrated gradients, and attention heatmaps, can reveal the internal decision logic of complex models such as deep neural networks. This is critical not only for model debugging and regulatory compliance, but also for building trust among system operators.

Ultimately, achieving robust and explainable machine learning systems for IoT security is not a one-and-done solution, but an ongoing process. It involves building resilience into every stage, from model training to deployment, while ensuring that explanations are accessible and actionable. This dual emphasis on security and transparency ensures that machine learning-driven IoT defenses are not only robust but also reliable and traceable in the face of real-world challenges

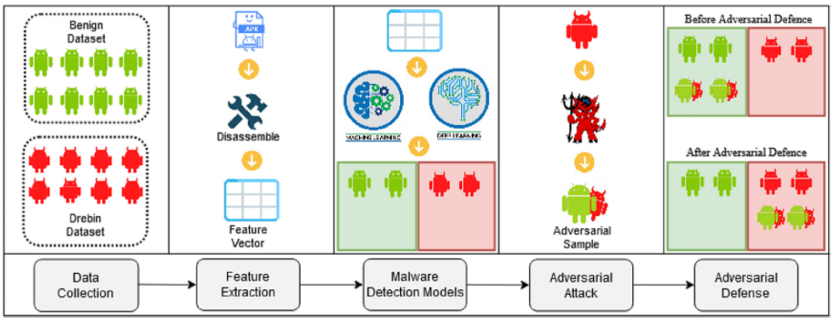

Fig. 2. Proposed framework for constructing robust malware detection model(s) using adversarial attack and defense strategies

5 Experiment on Machine Learning Models for IoT Traffic Detection

In an experimental comparison using the BoT-IoT dataset, deep learning-based Model A significantly outperformed traditional Model B in all key metrics. Confusion matrix analysis further confirmed Model A superior ability to detect malicious traffic and minimize false positives. Traditional models, while simpler and more resource-efficient, struggle with complex IoT traffic patterns. Deep learning models, while more computationally demanding, have better generalization, adaptability, and reliability, making them more suitable for practical IoT security applications.

5.1 Experimental Design

To evaluate the performance of machine learning models in IoT malicious traffic detection, we built a comprehensive and controlled experimental framework. The framework is designed to evaluate the effectiveness of two widely used modeling paradigms. Model A adopts advanced deep learning architectures such as convolutional neural networks (CNNs) and long short-term memory networks (LSTMs). These models are well suited to capture the complex spatiotemporal characteristics of network traffic patterns in dynamic IoT environments. CNNs excel at extracting local spatial features from traffic data, while LSTMs are able to learn long-term dependencies and sequences, which makes them particularly useful in time series classification tasks.

On the other hand, Model B represents traditional machine learning methods. Such algorithms include support vector machines (SVMs) and random forests (RFs), among others. Unlike deep learning models that automatically learn features, these models rely on manually designed features based on domain knowledge. Although they are less computationally intensive and easier to interpret, they may struggle to cope with the complexity and variability of IoT traffic data.

The experimental study uses the BoT-IoT dataset, a publicly available benchmark dataset containing a rich set of normal and attack traffic, including DDoS, DoS, reconnaissance, and information theft. The dataset was preprocessed to remove redundant and irrelevant features, normalize continuous values, and encode categorical variables. After preprocessing, the data was split into training and test sets in a 70:30 ratio to ensure that the evaluation metrics reflect the generalization ability of the model.

To ensure a fair comparison, Model A and Model B were trained under the same computing settings. The hyperparameters of each model were optimized using a grid search method, which exhaustively searches a set of predefined parameters to determine the best performing configuration. This tuning process is critical to maximize model performance and avoid underfitting or overfitting.

To evaluate the models objectively, four performance metrics were used. Accuracy measures the proportion of correctly classified instances to all instances. Precision calculates the ratio of true positive predictions to the total number of positive predictions, indicating the model's ability to minimize false positives. Recall determines the ratio of true positive predictions to the actual number of positive examples, reflecting the sensitivity of the model to detect attacks. Finally, the false positive rate (FPR) measures the proportion of benign traffic that is misclassified as malicious traffic, which is critical in the real world as false positives can overwhelm system administrators and reduce people's trust in intrusion detection systems (IDS).

The results of this experimental setup not only enable us to conduct a rigorous comparison between deep learning and traditional methods, but also help us gain insight into the trade-offs between computational complexity, detection performance, and practical deployment feasibility. The framework ensures reproducibility and can be extended to evaluate other machine learning models or datasets, laying a solid foundation for future research on intelligent IoT security systems.

5.2 Results and Discussion

This Fig. 3 provides a comprehensive and visual comparison of two machine learning models: Model A, which is based on a deep learning architecture, and Model B, which uses traditional machine learning methods. Model B was evaluated based on four core performance metrics: Accuracy, Precision, Recall, and False Positive Rate. These metrics are critical to evaluating the effectiveness and usefulness of intrusion detection systems in environments, where network integrity and operational continuity are critical.

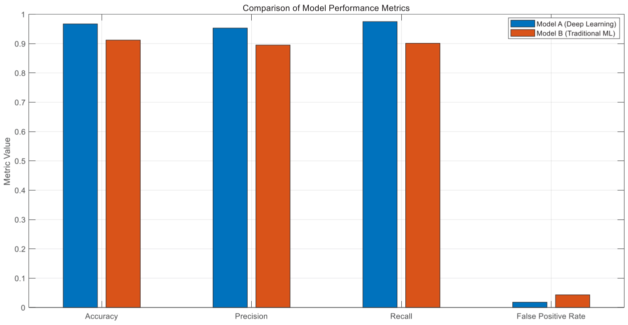

Fig. 3. Bar chart of Accuracy, Precision, Recall, FPR

Starting with accuracy, which represents the overall percentage of correctly classified instances, Model A significantly outperforms Model B. Model A has an accuracy of 97%, while Model B has an accuracy of only 91%. This result suggests that deep learning models are better able to capture the overall structure and variation of network traffic data, resulting in fewer misclassifications. In security systems, high accuracy is a strong indicator of model reliability and is critical to maintaining trust in automated decisions.

Accuracy, which refers to the ratio of a model's true predictions to all positive predictions, is another indicator that Model A clearly outperforms. Model A's accuracy is 95%, while Model B's is 90%, proving that Model A is more effective in reducing false positives. This is critical in operational environments, as too many false positives burden administrators and may cause them to overlook real threats.

Recall, which measures the model's ability to detect actual malicious traffic by calculating the proportion of true positive cases to the total number of actual positive cases, highlights another advantage of Model A. Its recall rate is 97.5%, significantly higher than Model B 90%. This difference means that Model A is more successful in detecting most malicious traffic, reducing the likelihood of undetected violations, which is particularly important in scenarios involving sensitive data or critical services.

False positive rate, a key metric in real-world deployments, shows a significant gap. This metric refers to the proportion of traffic that is incorrectly classified as malicious. Model A has a lower false positive rate of 1.8%, while Model B has a much higher false positive rate of 4.3%. A lower false positive rate is critical in IoT networks, as unnecessary blocking of legitimate traffic can lead to functional disruptions, communication delays, or a degraded user experience. Especially in critical environments such as smart grids or healthcare systems, minimizing false positives ensures stable and trustworthy system behavior.

These performance differences highlight the ability of deep learning models to identify and generalize complex spatiotemporal relationships in IoT traffic data. These models can automatically learn multi-level features directly from raw input without relying on human input, which improves their adaptability and scalability. However, deep learning models require a lot of computing resources, including more memory and processing power, which may limit their usability in resource-constrained IoT devices. To alleviate this problem, techniques such as model pruning, quantization, knowledge distillation, and edge-cloud integration can be applied to reduce model size and latency while maintaining accuracy.

In summary, the bar charts demonstrate the huge advantages that deep learning provides in IoT traffic detection across multiple metrics. While traditional models are still useful in lightweight applications, deep learning solutions provide higher detection performance, higher reliability, and better generalization, making them strong candidates for smart security systems in next-generation IoT deployments.

The Fig. 4 for Model A and Model B provide an in-depth assessment of their classification behavior in detecting IoT network traffic, especially in distinguishing between benign and malicious samples. These visualizations not only allow for comparative performance analysis, but also provide insights into how each model handles different categories, highlighting their strengths and weaknesses in actual operation. Taking the deep learning-based Model A as an example, the matrix shows that it correctly classified 930 benign instances and 1035 malicious instances. The total number of misclassifications is very low, with only 35 instances, of which 20 benign samples were incorrectly labeled as malicious and 15 malicious samples were missed. The normalized values further demonstrate its excellent performance: the recall rate for malicious traffic is 98.6% and the recall rate for benign traffic is 97.9%. This means that the model is highly sensitive to both attack and non-attack traffic, thereby minimizing threat exposure and minimizing interference with legitimate communications. Its accuracy is also very high, with 98.1% of malicious samples and 98.4% of benign samples being correctly classified. Additionally, the false positive rate was only 1.6%, which is critical to reducing unnecessary alerts that could otherwise cause alert fatigue or performance degradation in automated response systems.

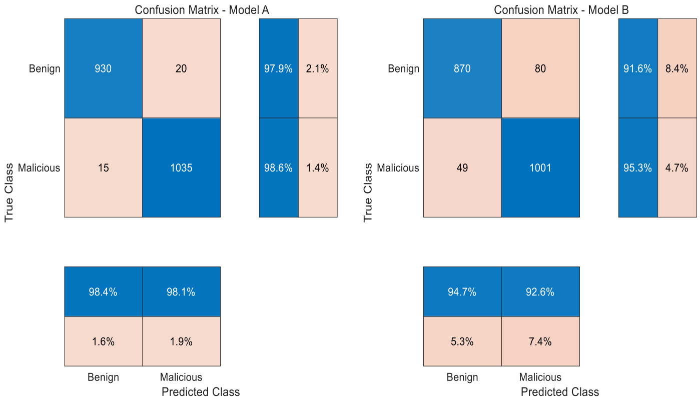

Fig. 4. Confusion matrices

In contrast, Model B, which represents traditional machine learning, performed significantly worse. It misclassified 80 benign samples as malicious, with a false positive rate of 8.4%. At the same time, it failed to detect 49 malicious instances, with a false positive rate of 4.7%. These flaws are particularly concerning in IoT environments, where undetected attacks can compromise device functionality or data integrity, while false positives can disrupt mission-critical services. The overall recall rate for benign traffic dropped to 91.6%, while the malicious recall rate remained high at 95.3%. However, its precision was lower than Model A, especially for benign samples, which means that the incidence of false positives was higher.

These differences highlight a key point: while traditional machine learning methods may work for simpler scenarios or situations where computing resources are highly constrained, they often struggle with the complexity and high dimensionality of IoT network traffic. They rely on hand-crafted features and may not capture subtle or evolving attack signatures. Model A, through its deep learning architecture, can use spatiotemporal learning to infer complex patterns that vary over time and space, thereby improving its adaptability and robustness to known and new threats.

In real-time applications, both detection accuracy and system reliability are critical, and these results strongly support Model A. Its lower false positive and false negative rates ensure fewer interruptions and enhance people's trust in the automated decision-making process. This is critical for deployment in smart homes, industrial control systems, medical IoT, and other critical infrastructure areas.

6 Conclusion

In summary, the experimental results of this study clearly demonstrate the effectiveness of machine learning for detecting malicious traffic in IoT environments. By using performance metrics and confusion matrices for comparative analysis, we verified that deep learning models significantly outperform traditional machine learning methods in terms of accuracy, precision, recall, and false alarm rate. However, while these results affirm the technical feasibility of intelligent detection systems, they also reveal several unresolved challenges that must be addressed for successful deployment in real-world applications. One of the most important issues is the reliance on outdated or overly simplified datasets that do not accurately reflect the ever-changing heterogeneous nature of real IoT traffic. This gap in data authenticity hinders the generalization ability of trained models and limits their performance in dynamic environments facing unknown threats. To improve the effectiveness of real-world applications, future research should prioritize the development and use of rich and continuously updated datasets that capture the operational diversity of IoT deployments. Another significant challenge lies in the inherent computational limitations of edge IoT devices. Despite their superior performance, deep learning models are resource-intensive, which limits their deployment on devices with limited processing power, memory, and energy availability. This highlights the urgent need for lightweight, optimized models that can reduce computational overhead while providing comparable detection accuracy. Integrating techniques such as model compression, quantization, and neural architecture search is essential to building viable edge computing models. In addition, the lack of explainability in most machine learning-based intrusion detection systems presents another obstacle. These systems often act as black boxes with little transparency into the decision-making process. In mission-critical applications such as healthcare and industrial automation, explainability is essential for debugging, regulatory compliance, and building stakeholder trust. Therefore, future work should focus on integrating explainable artificial intelligence (XAI) techniques that can provide meaningful insights into model behavior without compromising performance. Security and robustness are also pressing issues to be addressed. As attackers increasingly target machine learning models through evasion and poisoning attacks, building resilient detection mechanisms is critical. However, research on adversarial defenses remains relatively weak in the field of IoT. There is a growing need for comprehensive approaches that combine robustness with transparency to ensure national defense strength and operational reliability. Finally, the lack of standardized evaluation protocols and benchmarking frameworks hinders reproducibility and objective comparison between studies. Establishing shared benchmarks covering diverse datasets, model evaluation metrics, and deployment scenarios would significantly advance the field and promote more systematic innovation.

7 Acknowledgement

Project Name: Research on Malicious Traffic Identification in the Internet of Things Based on Deep Learning
Xingtai City's Municipal Science and Technology Plan Self - funded Project (2024ZC040).

References

[1] A.-J. Pinheiro, J. de M. Bezerra, C.-A.-P. Burgardt, D.-R. Campelo, Identifying IoT devices and events based on packet length from encrypted traffic, Computer Communications 144(2019) 8-17.

[2] S.-Z. Zhu, X.-L. Xu, H.-H. Gao, F. Xiao, CMTSNN: A Deep Learning Model for Multiclassification of Abnormal and Encrypted Traffic of Internet of Things, IEEE Internet of Things Journal 10(13)(2023) 11773-11791.

[3] T.-D. Diwan, S. Choubey, H.-S. Hota, S.-B. Goyal, S.-S. Jamal, P.-K. Shukla, B. Tiwari, Feature Entropy Estimation (FEE) for Malicious IoT Traffic and Detection Using Machine Learning, Mobile Information Systems 2021(2021) 8091363.1-8091363.13.

[4] H. Kawai, S. Ata, N. Nakamura, I. Oka, Identification of communication devices from analysis of traffic patterns, in: Proc. 2017 13th International Conference on Network and Service Management, 2017.

[5] F.-B. Islam, C.-I. Nwakanma, J.-M. Lee, D.-S. Kim, Enhancing Malicious Activity Classification of IoT Network Traffic Characteristics using Stacked Ensemble Learning, in: Proc. 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), 2021.

[6] N. Koroniotis, N. Moustafa, E. Sitnikova, B. Turnbull, Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset, Future Generation Computer Systems 100(2019) 779-796.

[7] N.-D. Liao, J.-Y. Guan, Multi-scale Convolutional Feature Fusion Network Based on Attention Mechanism for IoT Traffic Classification, International Journal of Computational Intelligence Systems 17(1)(2024) 36.1-36.25.

[8] A. Hamza, H.-H. Gharakheili, T.-A. Benson, V. Sivaraman, Detecting Volumetric Attacks on IoT Devices Via SDN-Based Monitoring of MUD Activity, in: Proc. of the 2019 ACM Symposium on SDN Research, 2019.

[9] I. Yaqoob, E. Ahmed, M.-H. Rehman, A.-I.-A. Ahmed, M.-A. Al-garadi, M. Imran, M. Guizani, The rise of ransomware and emerging security challenges in the Internet of Things, Computer Networks 129(2017) 444-458.

[10] Z. Abou El Houda, A. Hafid, L. Khoukhi, Co-IoT: A Collaborative DDoS mitigation scheme in IoT environment based on blockchain using SDN, in: Proc. IEEE Global Communications Conference, 2019.

[11] D. M. Sharif, H. Beitollahi, Detection of application-layer DDoS attacks using machine learning and genetic algorithms, Computers & Security 135(2023) 103511.1-103511.5.

[12] A. Alabdulatif, I. Khalil, M.-S. Rahman, Security of Blockchain and AI-Empowered Smart Healthcare: Application-Based Analysis, Applied Sciences 12(21)(2022) 11039.1-11039.32.

[13] S. Ali, O. Abusabha, F. Ali, M. Imran, T. Abuhmed, Effective Multitask Deep Learning for IoT Malware Detection and Identification Using Behavioral Traffic Analysis, IEEE Transactions on Network and Service Management 20(2)(2023) 1199-1209.

[14] C.-P. Fu, Q. Li, K. Xu, Flow Interaction Graph Analysis: Unknown Encrypted Malicious Traffic Detection, IEEE/ACM Transactions on Networking 32(4)(2024) 2972-2987.

[15] Y.-B. Feng, J. Li, L. Jiao, X.-T. Wu, Towards Learning-Based, Content-Agnostic Detection of Social Bot Traffic, IEEE Transactions on Dependable and Secure Computing 18(5)(2021) 2149-2163.

[16] E. Papadogiannaki, S. Ioannidis, A Survey on Encrypted Network Traffic Analysis Applications, Techniques, and Countermeasures, ACM Computing Surveys 54(6)(2022) 1-35.

[17] S.-I. Imtiaz, L.A. Khan, A.S. Almadhor, S. Abbas, S. Alsubai, M. Gregus, Z. Jalil, Efficient Approach for Anomaly Detection in Internet of Things Traffic Using Deep Learning, Wireless Communications and Mobile Computing, 2022(1)(2022) 8266347.1-8266347.15.

[18] A. Rahim, Y.-R. Zhong, T. Ahmad, S. Ahmad, P. Pławiak, M. Hammad, Enhancing Smart Home Security: Anomaly Detection and Face Recognition in Smart Home IoT Devices Using Logit-boosted CNN Models, Sensors 23(15)(2023) 6979.1-6979.42.

[19] S.-M. Shagari, D. Gabi, N.-M. Dankolo, N.-N. Gana, Countermeasure to Structured Query Language Injection Attack for Web Applications Using Hybrid Logistic Regression Technique, Journal of the Nigerian Society of Physical Sciences 4(4)(2022) 832.1-832.15.

[20] R. Jahangir, Y.-W. Teh, N.-A. Memon, G. Mujtaba, M. Zareei, U. Ishtiaq, Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network, IEEE Access 8(2020) 32187-32202.

[21] K. Çoşkun, G. Çetin, A Comparative Evaluation of the Boosting Algorithms for Network Attack Classification, International Journal of 3D Printing Technologies and Digital Industry 6(1)(2022) 102-112.

[22] S.-G. Liu, Y. Wang, J. Zhang, C. Chen, Y. Xiang, Addressing the class imbalance problem in Twitter spam detection using ensemble learning, Computers & Security 69(2017) 35-49.

[23] N. Koroniotis, N. Moustafa, E. Sitnikova, B. Turnbull, Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset, Future Generation Computer Systems 100(2019) 779-796.

[24] C. Wheelus, E. Bou-Harb, X.-Q. Zhu, Tackling Class Imbalance in Cyber Security Datasets, in: Proc. IEEE International Conference on Information Reuse and Integration, 2018.

[25] A.-S. Barkah, S.-R. Selamat, Z.-Z. Abidin, R. Wahyudi, Data Generative Model to Detect the Anomalies for IDS Imbalance CICIDS2017 Dataset, TEM Journal 12(1)(2023) 80-89.

[26] M.-N. Rahman, A. Esmailpour, J.-H. Zhao, Machine Learning with Big Data: An Efficient Electricity Generation

Forecasting System, Big Data Research 5(2016) 9-15.

[27] C.-A. Keller, M.-J. Evans, J.-N. Kutz, S. Pawson, Machine learning and air quality modeling, in: Proc. IEEE International Conference on Big Data, 2017.

[28] V.-E. Sathishkumar, J. Cho, M. Subramanian, O.-S. Naren, Forest fire and smoke detection using deep learning-based learning without forgetting, Fire Ecology 19(2023) 9.1-9.17.